# Updates to PDF indexing for search results

Last Modified on 12/28/2022 12:24 pm EST

Today, I'd like to tell you a story about PDFs and how they're indexed for search.

When you add a PDF to an article--whether as a link to the PDF or in an iframe--once you save the article, we run a background job that scrapes the text out of the PDF so we can include it in the index of that article for search. By default, we do this for PDFs under 100 pages only, but there is an option in **Settings > Search** so that you can turn it on for PDFs over 100 pages.

This process is not an exact science. If you've ever seen search results with the PDF logo and some weird characters next to them, it's the PDF scraping/parsing to blame.

However, this story has a somewhat happy ending (for now): one of our developers spent some quality time with the PDF parser over the holidays, and we've just released some exciting improvements to it.

## Improvement #1

Which is basically a bug fix:

If you previously uploaded long PDFs into articles, and then checked the box to index PDFs over 100 pages, long PDFs within iframes would not show in search results.

We've updated the indexing process here so that those PDFs are properly indexed for search once the box is checked.

**If you are using this setting and have noticed PDFs in iframes weren't showing in search results, please contact us to request a search reindex of your knowledge base.**

- The reindex will pause search while it's running (usually under 5 minutes). Afterward you should see these PDFs in iframes properly appearing in search result blurbs.

## Improvement #2

We've improved the scraping portion of the PDF parsing so that it handles special characters better. Now we should be better handling things like:

- Accented characters or other non-English characters
- Bullet points and other unusual punctuation marks
- Non-breaking spaces

This is still an inexact science, but the changes we released yesterday do better handle those characters in most PDFs.

If you've noticed issues with PDF content showing up with odd characters or formatting in one or two search result article blurbs, you can make a small edit to the articles in question and re-save. That will trigger the new indexing process.

If you've noticed these issues with a lot of PDFs/articles, **please contact us to request a search reindex of your knowledge base.**

- The reindex will pause search while it's running (usually under 5 minutes). Afterward you should see these PDFs in iframes properly appearing in search result blurbs.